Aya Model: An Instruction Finetuned Open-Access Multilingual Language Model

Year: 2024 | Citations: 306 | Authors: A. Ustun, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza

Abstract

Recent breakthroughs in large language models (LLMs) have centered around a handful of data-rich languages. What does it take to broaden access to breakthroughs beyond first-class citizen languages? Our work introduces Aya, a massively multilingual generative language model that follows instructions in 101 languages of which over 50% are considered as lower-resourced. Aya outperforms mT0 and BLOOMZ on the majority of tasks while covering double the number of languages. We introduce extensive new evaluation suites that broaden the state-of-art for multilingual eval across 99 languages -- including discriminative and generative tasks, human evaluation, and simulated win rates that cover both held-out tasks and in-distribution performance. Furthermore, we conduct detailed investigations on the optimal finetuning mixture composition, data pruning, as well as the toxicity, bias, and safety of our models. We open-source our instruction datasets and our model at https://hf.co/CohereForAI/aya-101